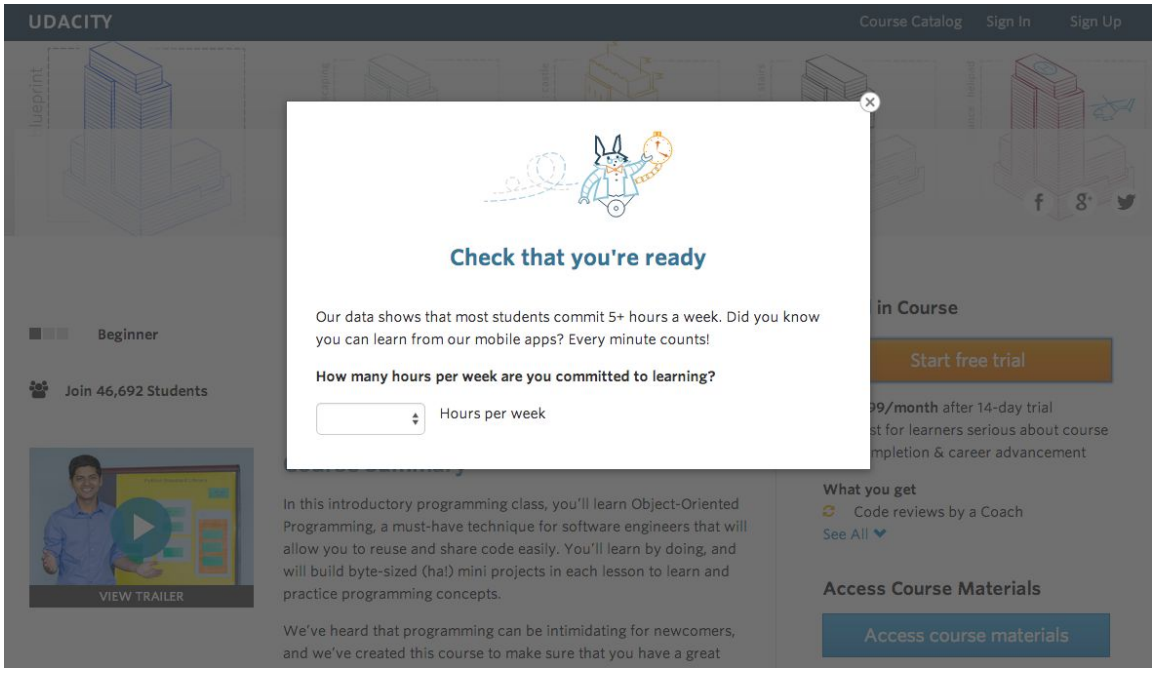
Design An A/B Test (Free Trial Screener)

Experiment Overview:

At the time of this experiment, Udacity courses currently have two options on the course overview page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time. The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The screenshot of the experiment is shown below:



Experiment Design

Metric Choice

Invariant Metrics: Invariant metrics are those metrics which should not change across the experimental and control groups. These metrics can play a good role in sanity checking. As part of this experiment, the following can be chosen as the invariant metrics.

- **Number of cookies (That is, number of unique cookies to view the course overview page.):** No of cookies is a good population sizing metric. As unit of diversion is cookie for this experiment, they should be randomly distributed across the control and experiment groups and both the groups should have approximately equal no of them to pass the initial sanity test.

- **Number of clicks (That is, number of unique cookies to click the "Start free trial" button, which happens before the free trial screener is trigger.) :** As the changes are made to the flow post the click on “start free trial” there should be no change in this parameter due to experiment and this parameter should be comparable across both the groups.

Evaluation Metrics: Evaluation metrics are those metrics which vary across the experimental and control groups, changes shown in these metrics are due to the changes in the system which are made for the experiment.

- **Gross conversion (That is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button)** : Due to the change in the flow of “start free trial”, students who can spend less than five hours per week might not go through the regular flow and might opt for the option “access course materials” in experiment group, which would decrease the numerator of this metric, thus decreasing the overall metric when compared to the control group.

- **Net conversion (That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button.):** In the experiment group, students who can spend more than five hours might complete the checkout process and stay enrolled after the 14 day boundary period, so the numerator of this metric might remain constant/decrease in the experiment group which would decrease the overall metric in experiment group, when compared to the control group.

Unused Metrics: These are metrics which will not be captured for the experiment.

- **Click-through-probability (That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.):** Click through probability is defined as the ratio of no of clicks to the no of pageviews, since both the parameters will be captured as part of above two invariant metrics, there would be no need explicitly capture this metric.

- **Retention (That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout.)** : Since this metric is a combination of other two evaluation metrics, there would be no need to explicitly capture this metric and unit of analysis of this metric is different from unit of diversion which would make our analytical estimate an underestimate and would require us to calculate this metric empirically.

In order to launch the experiment, the gross conversion metric should show a decrement in the experiment group when compared to control group and Net Conversion metric should either remain constant or should increase because the decrement in net conversion metric will impact revenue of the company.

Measuring Standard Deviation

Since both the chosen evaluation metrics involve probabilities, the standard deviation of these metrics can be calculated using the formula of standard deviation for binomial distribution. This spreadsheet contains rough estimates of the baseline values for these metrics. The baseline values mentioned the sheet correspond to 40,000 pageviews, but we need to estimate the standard deviation for the chosen evaluation metrics which correspond to 5,000 pageviews.

S No.	Evaluation Metric	For 40k Pageviews	For 5k Pageviews
1.	Gross conversion	0.007152	0.0202
2.	Net Conversion	0.005515	0.0156

Since the unit of diversion and unit of analysis of the chosen evaluation metrics is same, both the empirical and analytical estimates will be comparable.

Sizing

Number of Samples vs. Power

Bonferroni correction is not used during the analysis phase because the evaluation metrics are correlated and using it will be too conservative for our experiment as we have only two metrics. The number of pageviews required to power the experiment can be calculated using the Evan Miller's Sample Size Calculator. The baseline conversion rate and other required details for both the metrics can be captured from this spreadsheet.

S.No	Evaluation Metric	Samples Required (w.r.t Table Of Baseline Values)	Total Pageviews (For Both Groups In Exp)
1.	Gross conversion base Conversion Rate=20.625 (d_{min}= 0.01) alpha = 0.05 beta = 0.2	25,835	6,45,875 ((25,835/0.08)*2)
2.	Net Conversion base Conversion Rate=10.93125 (d_{min}= 0.075) alpha = 0.05 beta = 0.2	27,413	6,85,325 ((27,413/0.08)*2)

The total number of pageviews required to power the experiment is 6,85,325.

Duration vs. Exposure

As per the table of baseline values, the number of daily unique pageviews is 40,000, so if we decide to **divert 70%** of the daily traffic through our experiment, it would take around **25 days** to properly power the experiment.

We should not divert 100% of traffic through our experiment, because if there is a bug, it would affect entire user base.

As we know that weekends and weekdays activity is different, we might want to run the experiment on a series of weekends and weekdays, so that we acquire proper results.

We can consider this experiment as somewhat risky because it's impacting the regular checkout flow, any bug can result in revenue loss for the organisation.

Experiment Analysis

Sanity Checks

If the invariant metric is a simple count than it should be randomly split b/w the two groups. We can use a binomial test to determine if the observed value is within the expected range. Below is the data for 95% confidence interval of invariant metrics. The Data required to conduct the sanity test is available in [this spreadsheet].

S.No	Invariant Metric	Observed Val	Lower Bound	Upper Bound	Sanity Passed
1.	Number of cookies	0.5006	0.4988	0.5012	Yes
2.	Number of Clicks	0.5005	0.4959	0.5041	Yes

Result Analysis

Effect Size Tests

Below is the table of 95% confidence intervals for evaluation metrics. For a metric to be statistically significant its confidence interval should not contain zero and for a metric to be practically significant its (d_{min}) should lie beyond the confidence interval values.

S.No	Evaluation Metric	Confidence Interval	Statistically Sig.	Practically Sig.
1.	Gross Conversion **Control Group :** Clicks : 17293 Enrollments : 3785 **Experiment Group :** Clicks : 17260 Enrollments : 3423	[-0.02912,-0.01198] D_{Min} = 0.01 Observed Diff : - 0.02055 Pooled Pb : 0.20860 SD : 0.004372 ME : 0.008568	Yes	Yes
2.	Net Conversion **Control Group :** Clicks : 17293 Payments : 2033 **Experiment Group :** Clicks : 17260	[-0.01160,0.001857] D_{Min} = 0.0075 Observed Diff : -0.004874 Pooled Pb : 0.1151 SD : 0.003434 ME : 0.006731	No (Contains Zero)	No

	Payments : 1945			

Sign Tests

The effect size can also be calculated using non parametric sign test. The results can be used to establish whether a metric is statistically significant or not.

S.No	Evaluation Metric	P- Value	Statistically Sig
1.	Gross Conversion	0.0026 (two-tail P value) Number of "successes": 19 Number of trials: 23	Yes
2.	Net Conversion	0.6776 (two-tail P value) Number of "successes": 13 Number of trials: 23	No

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

Recommendation

Make a recommendation and briefly describe your reasoning.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.